



**Tecnológico
de Monterrey**

**Aplicación de métodos multivariados
en ciencia de datos (MA2003B.201)**

Etapa 4. Informe final del Reto

Sergio Alejandro Esparza González A01625430

Pedro Soto Juárez A00837560

Mauricio Octavio Valencia Gonzalez A01234397

Luis Fernando Alcázar Díaz A00836287

Docente:

Blanca Rosa Ruiz Hernandez

Socio Formador:

SIMA

11 de octubre de 2024

1. Resumen

El primer paso en este análisis consistió en revisar y consolidar todos los conjuntos de datos (datasets) correspondientes a los años 2022 a 2024. La razón principal para seleccionar estos años es que proporcionan una visión más actual y relevante de lo que está ocurriendo en la realidad, permitiendo un análisis con datos más frescos y representativos. A partir de aquí, se generaron diferentes dataframes de acuerdo a los diferentes usos que les dimos a los datos para lograr los diferentes objetivos que nos propusimos.

Se hicieron algunos procesos estadísticos en la base consolidada previo a dividirla, por ejemplo la conversión de outliers en valores nulos y la imputación de datos. Para hacer la imputación de datos se usó el algoritmo de imputación MICE (Multiple Imputation by Chained Equations) en combinación con la regresión bayesiana. Este método es más eficiente que las imputaciones simples, ya que considera la variabilidad entre las variables e incorpora incertidumbre en el modelo de regresión, lo que lo hace más adecuado para datasets con alta variabilidad o donde se busca evitar el sobreajuste. Además de esto se estandarizaron los datos, lo que significa que se ajustaron las variables para que tuvieran la misma escala, facilitando el análisis de patrones y la aplicación de algoritmos de clustering. (Azur, 2011)

Después de esto, se crearon las diferentes bases de datos; la del análisis descriptivo de las variables, la de los histogramas de datos sin procesar y los procesados, la base con los datos estandarizados, la base utilizada para la clusterización, la de la interpolación y por último una base de datos para la ubicación de las estaciones de monitoreo.

Ya teniendo las bases de datos limpias y procesadas se hizo un análisis de componentes principales, para los diferentes contaminantes e industrias analizadas. Se acotaron los datos al 25% de estaciones de monitoreo más cercanas a las fuentes de contaminación y a los contaminantes relevantes, que fueron identificados con los análisis de las variables e histogramas previamente mencionados así como gráficas de distribución e investigación en la literatura recomendada por el socio formador, así como literatura relevante externa. El segundo método utilizado fue el de KNN Clusters, para el cual se agrupó en las estaciones más afectadas, los contaminantes que más las afectan. Los resultados obtenidos en ambos procesos fueron similares, se pudo observar una diferencia significativa entre la concentración de los contaminantes en las diferentes estaciones.

Por último, se hizo una interpolación de datos de acuerdo a la ubicación de las estaciones de monitoreo, la concentración de los contaminantes, la dirección y velocidad del viento. Se uso un metodo de interpolacion cubica y para representar los datos, se hizo la interpolación para las diferentes horas del día y que de esta forma se notara el cambio en la contaminación de la ciudad y cómo evoluciona a lo largo de toda la ciudad y no solo en los puntos de monitoreo.

2. Introducción y justificación

Uno de los problemas más constantes, y sin embargo, más ignorados es el de la mala calidad del aire. Respirar aire contaminado se ha aceptado como un aspecto negativo pero inevitable de la vida moderna, especialmente en ciudades industriales como Monterrey; un pequeño precio a pagar por el desarrollo económico, pero esta resignación nos está costando nuestra salud y nuestra vida. La Organización Mundial de la Salud (2019) estima que el 99% de la población mundial respira aire que excede los límites establecidos por la organización.

Es por esto que para atacar esta problemática en Monterrey, la medición continua de la calidad del aire dio inicio al Sistema Integral de Monitoreo Integral (SIMA), con el objetivo de tener un sistema que tenga información fidedigna y actualizada de niveles de contaminación ambiental en la Zona Metropolitana de Monterrey (ZMM), se busca conocer el impacto de la contaminación sobre los sectores poblacionales a partir de estudios metereológicos, epidemiológicos y de uso de suelo.

Es por esto que medir la calidad del aire resulta tan crucial para conocer este impacto, pues lo que no se mide no se mejora y al obtener los datos y analizarlos podrían incluso salir a la luz nuevas problemáticas que no se tenían contempladas previamente. Los contaminantes monitoreados en estaciones tales como Santa Catarina, San Bernabé, Obispado, San Nicolás, La Pastora, son Monóxido de carbono (CO), partículas menores a 10 micras (PM10), ozono (O₃), óxidos de nitrógeno (NO, NO₂) y dióxido de azufre (SO₂), siendo estos considerados internacionalmente, los principales contaminantes en el aire.

Medir la calidad del aire enfrenta varios obstáculos, entre ellos la dificultad para medir las emisiones de fuentes difusas y la complejidad de la composición química del material particulado, lo que dificulta evaluar el impacto en la salud, se requiere una integración de fuentes interiores y exteriores. Los avances en sensores de bajo costo deben garantizar la calidad de los datos. (Sokhi et al., 2022)

Pero a pesar del avance en la tecnología de los sensores, inevitablemente habrá “puntos ciegos” entre los sensores, en donde no se tienen datos, lo único que se puede lograr al aumentar la cantidad de sensores es disminuir el tamaño de estos puntos ciegos, más no eliminarlos. Para lograr eliminar estos puntos ciegos, se pueden hacer estimaciones de los valores de acuerdo a las concentraciones de los contaminantes en estaciones cercanas, así como datos meteorológicos que pueden afectar la concentración en los puntos intermedios, como la velocidad del viento.

Se quiere dar visibilidad al impacto de actores claros y reconocibles en la ciudad, para que desde esas fuentes principales se pueda atacar la contaminación generada y se puedan implementar soluciones, con un impacto mayor en la contaminación urbana. Además, al obtener mapas que puedan brindar información relevante sobre la concentración de la contaminación en toda la ciudad se puede poner a disposición de la ciudadanía una herramienta importante sobre la contaminación geolocalizada y qué significa esto para la salud de los ciudadanos, para que así, todos puedan tomar medidas correspondientes para

minimizar el efecto de los contaminantes y a largo plazo, reducir la presencia de estos en la atmósfera.

3. Problemática y Objetivo

El primer objetivo de este proyecto fue investigar el fenómeno del smog en la ciudad. Se planteó analizar los diferentes contaminantes que lo componen con la intención de desarrollar un modelo predictivo que pudiera anticipar su presencia en la ciudad durante los siguientes cuatro días, utilizando como base los últimos datos registrados. Sin embargo, durante el proceso de investigación, se descubrió que los niveles de los contaminantes asociados al smog rara vez excedían los límites aceptables establecidos por las normas de calidad del aire.

Debido a este hallazgo, el enfoque se re-ajustó para centrarse en los compuestos que tienen un impacto más significativo en la contaminación ambiental de la región. A partir del análisis de los datos, se identificaron los principales contaminantes que afectan la calidad del aire en Monterrey que son: el PM2.5, PM10, NO_x, SO₂ y CO, que además, son compuestos con un porcentaje alto de su emisión proveniente de la industria, lo que llevó a reorientar el estudio hacia la contaminación industrial.

Como ya se mencionó antes, se estima que el 99% de la población mundial respira aire que excede los límites establecidos por la organización. De estos, los contaminantes que se consideran más dañinos para la salud son las partículas en suspensión (PM2.5 y PM10) ya que estas pueden penetrar profundamente en los pulmones y el sistema cardiovascular, causando enfermedades respiratorias y cardiovasculares. Además, el Dióxido de Nitrógeno (NO₂) se asocia con enfermedades respiratorias y puede disminuir la función pulmonar. El Dióxido de Azufre (SO₂) puede formar lluvia ácida y dañar a las plantas y estructuras de construcción. Y el Monóxido de Carbono (CO) puede interferir con la capacidad de la sangre para transportar oxígeno, afectando el sistema cardiovascular. (Universidad de Buenos Aires, 2008)

Por lo tanto, gracias a la identificación de esta problemática, el objetivo cambió a determinar cuáles son los contaminantes predominantes en la atmósfera, considerando su distribución geográfica, y establecer una correlación con la ubicación de las plantas industriales más relevantes dentro de la Zona Metropolitana de Monterrey. Monterrey es una región con alta actividad industrial y tráfico vehicular, lo que contribuye a la emisión de estos contaminantes clave.

La selección de esta nueva línea de investigación se llevó a cabo con base en los datos proporcionados por el Programa Integral de Gestión de la Calidad del Aire (PIGECA) y las discusiones sostenidas con el socio formador. Esta información permitió identificar las áreas de mayor interés y establecer un enfoque estratégico para entender la relación entre la contaminación industrial y su impacto en la región, evaluando cómo la proximidad a las principales fuentes de emisión influye en la calidad del aire local.

Se toman en cuenta factores como dirección del viento, velocidad y temperatura. Los objetivos de la investigación de la dirección del aire es analizar y resumir ciertos tiempos de

mediciones de las concentraciones de los contaminantes, documentando características de regiones de Monterrey por medio de concentraciones hora-hora. (Centro de Calidad Ambiental, 2015)

4. Técnicas estadísticas

Esto fue realizado en el código: [VERIFICACION.Rmd](#)

4.1. Relaciones de dependencia

Método 1

Pruebas	Valor	Interpretación
Breusch-Pagan Test	p < 0.04	Existe heterocedasticidad, lo que significa que la variabilidad de los errores no es constante.
Durbin Watson Test	p < 2.2e^-16	Rechazamos la hipótesis nula (H_0). Esto nos lleva a aceptar que existe autocorrelación positiva en los residuos del modelo.
Multicolinealidad (VIF)	NOx, PM10, PM2.5 y SO2 < 10	Dado que los valores de VIF son menores a 10, no rechazamos H_0 , lo que indica que no hay un problema significativo de multicolinealidad.

Tabla 1. Pruebas estadísticas.

Estos resultados sugieren que, aunque el modelo enfrenta desafíos importantes, como la heterocedasticidad y la autocorrelación positiva, estas condiciones son comprensibles debido al uso de datos que abarcan varios años. La variabilidad de los errores (heterocedasticidad) y los patrones en los residuos (autocorrelación) suelen ser comunes en análisis de datos históricos, ya que los cambios en el tiempo pueden influir en la estructura de los errores.

La baja multicolinealidad es un aspecto positivo, ya que garantiza que las variables explicativas sean lo suficientemente independientes entre sí, lo que evita redundancias. La estructura de las variables independientes se considera adecuada, lo cual contribuye a la fiabilidad general del modelo.

Método 2. Regresión múltiple

También se hizo un modelo de regresión múltiple para profundizar un poco más sobre los supuestos:

```
lm(CO_ajustado ~ NOX + PM10 + PM2.5 + SO2, data =  
df_completo_estandarizado_2022_2024)
```

1. Transformación de Box-Cox: Se aplicó para estabilizar la varianza de la variable dependiente CO. Se seleccionó un valor óptimo de lambda para mejorar el ajuste del modelo.
2. Modelo de Regresión Múltiple:
 - Variables: Se ajustó un modelo con CO_transformado como variable dependiente y NOX, PM10, PM2.5 y SO2 como independientes.
 - Resultados: Ninguna de las variables predictoras mostró un efecto significativo en CO, con p-values altos (todos > 0.05). Esto sugiere que no están explicando bien la variabilidad de CO.
3. Modelo Corregido (Cochrane-Orcutt):
 - Se corrigió el modelo por autocorrelación, pero se advirtió que el ajuste es casi perfecto, lo que puede indicar sobreajuste.
 - R² negativo sugiere problemas en el ajuste.
4. Diagnóstico del Modelo:
 - Heterocedasticidad: La prueba de Breusch-Pagan indica variabilidad no constante en los residuos (p-value < 0.05).
 - Autocorrelación: El estadístico Durbin-Watson sugiere autocorrelación en los residuos, lo que es problemático para la validez del modelo.
 - Colinealidad: Los valores de VIF son bajos, indicando que no hay problemas serios de colinealidad entre las variables.

Conclusiones

- Limitaciones: Las variables no son significativas y el modelo presenta problemas de ajuste (heterocedasticidad y autocorrelación).
- Recomendaciones: Revisar los datos y considerar otras variables o transformaciones para mejorar el modelo.

4.2. Relaciones de interdependencia

Para modelar las relaciones de interdependencia se utilizaron lo siguiente. KNN, con 4 clusters y análisis PCA, los resultados de estos métodos se presentarán en la sección análisis, puesto que son los resultados con los que se decidieron trabajar.

Método 1. KNN

Presente en: [SCRIPT4_CLUSTERS.ipynb](#)

Para evaluar la selección de las estaciones de monitoreo de calidad del aire, utilizamos el algoritmo de KNN (K-Nearest Neighbors). Este método permitió agrupar las plantas industriales en función de los contaminantes que las afectan más significativamente. Los contaminantes analizados fueron CO, NOX, PM10, PM2.5, y SO2.

Método 2. PCA

Presente en: [SCRIPT5_GRAFICACIONVIENTO.ipynb](#)

Elegimos el Análisis de Componentes Principales porque nos ayuda a reducir la cantidad de variables sin perder información importante. Esto hace que sea más fácil analizar los contaminantes, ya que nos permite identificar patrones clave y eliminar cosas que se repiten, lo cual hace que el análisis sea más eficiente.

4.3. Métodos fallidos

Se llevó a cabo una regresión lineal para analizar el comportamiento de las variables, específicamente de NOx. Los resultados fueron extremadamente favorables, alcanzando una precisión de aproximadamente 99% en la predicción. Sin embargo, esta cifra se consideró poco realista, lo que llevó a la conclusión de que sería fantasioso confiar en la capacidad de la regresión lineal para proporcionar tal exactitud. Por esta razón, se decidió no realizar análisis adicionales basados en este modelo.

CFE

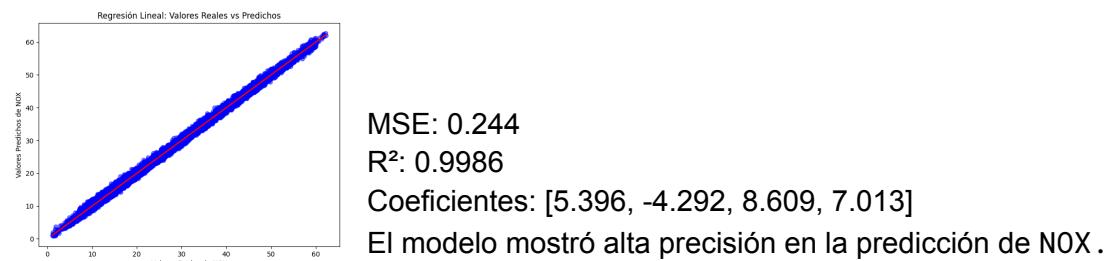


Imagen 1. Regresión lineal de NOX para planta CFE.

TERNIUM

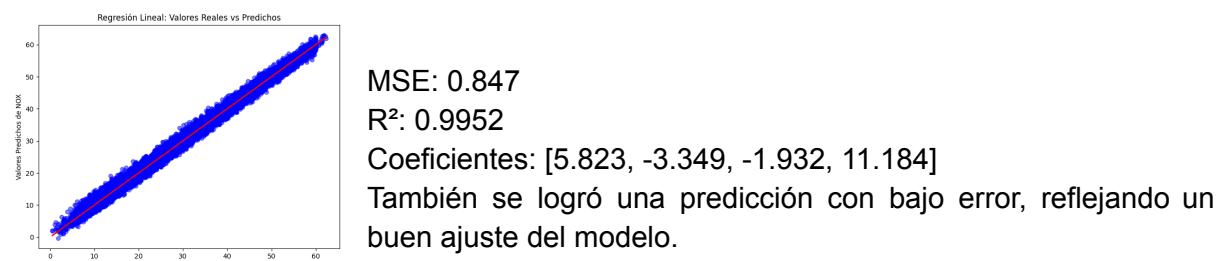
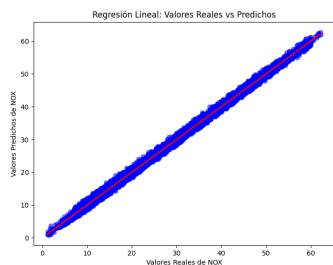


Imagen 2. Regresión lineal de NOX para planta TERNIUM.

CEMEX



MSE: 0.244

R²: 0.9986

Coeficientes: [5.396, -4.292, 8.609, 7.013]

Similar a CFE, el modelo es altamente preciso en sus predicciones de NOx

Imagen 3. Regresión lineal de NOX para planta CEMEX.

5. Contexto estaciones

Es importante describir el contexto de la estación metereológica que analizaron (ubicación y factores geográficos que podrían influir en la recopilación de las medidas meteorológicas y contaminantes)

Tomamos nuestras decisiones basadas en una revisión exhaustiva de la bibliografía, enfocándonos en identificar las empresas más relevantes dentro de las áreas específicas de interés. Utilizamos como fuente principal el documento de [PIGECA_2023_2033-1.pdf](#), lo cual nos permitió identificar las industrias que generan la mayor cantidad de contaminación en relación con los contaminantes que teníamos en nuestro conjunto de datos. Además, obtuvimos las coordenadas geográficas de las plantas industriales de estas empresas para extender el análisis visual.

Tabla 13. Porcentaje de participación por sector en las emisiones totales del inventario del año 2018

Sector	Emisiones de contaminantes año 2018 (t/año)						
	PM ₁₀	PM _{2.5}	SO ₂	CO	NO _x	COV	NH ₃
Fuentes fijas	59%	58%	99%	4%	21%	10%	3%
Generación de energía eléctrica	10.0%	14.5%	0.1%	1.9%	8.2%	0.2%	0.9%
Metalúrgica (incluye la siderúrgica)	9.4%	10.4%	0.2%	0.5%	1.9%	0.4%	0.6%
Petróleo y petroquímica	8.0%	6.4%	93.5%	0.4%	2.8%	0.8%	0.5%
Minerales no metálicos	7.3%	6.8%	0.1%	0.1%	0.3%	0.1%	0.1%
Extracción/Beneficio minerales no metálicos	7.2%	4.4%	0.0%	0.0%	0.1%	0.0%	0.0%
Automotriz	6.6%	4.7%	0.0%	0.3%	0.4%	2.0%	0.1%
Vidrio	2.4%	2.4%	1.0%	0.2%	2.7%	0.2%	0.2%
Química	2.0%	1.9%	0.3%	0.1%	1.8%	2.2%	0.1%
Cemento y cal	1.9%	1.7%	3.8%	0.6%	1.8%	0.0%	0.1%
Accesorios, aparatos eléctricos y equipos de generación eléctrica	1.1%	1.3%	0.0%	0.0%	0.1%	0.5%	0.0%
Metálico	0.8%	0.9%	0.0%	0.1%	0.2%	0.6%	0.1%
Alimentos y bebidas	0.7%	0.4%	0.3%	0.0%	0.1%	0.0%	0.0%
Plástico y hule	0.5%	0.4%	0.0%	0.0%	0.0%	0.6%	0.0%
Celulosa y papel	0.4%	0.5%	0.0%	0.2%	0.4%	0.1%	0.1%

Tabla 2. Porcentaje de participación por sector en las emisiones totales del inventario del año 2018.

Aquí están las principales actividades industriales de interés:

Generación de energía eléctrica: Este sector es responsable de grandes emisiones de NOx, CO y PM10.

Empresas clave:

- CFE (varias subestaciones):
 - Subestación Tecnológico
 - Subestación Mirador
 - Subestación San Agustín
 - Subestación San Jerónimo Potencia
 - Subestación Obispado
 - Subestación Nogalar

Metalurgia: Este sector también contribuye significativamente a las emisiones de NOx, CO y PM10.

Empresas clave:

- Ternium:
 - Ternium 1
 - Lagos NL
 - Ternium 2
 - Ternium Churubusco
 - Ternium CEDI
 - Ternium Guerrero

Producción de cemento y cal: Se asocian con emisiones elevadas de NOx, CO, PM10 y SO2.

Empresas clave:

- Cemex:
 - Planta Cemex
 - Planta Cemex Morones
 - Planta Cemex Cumbres

Las ubicaciones con coordenadas se encuentran aquí: [Ubicacion_Industrias.csv](#)

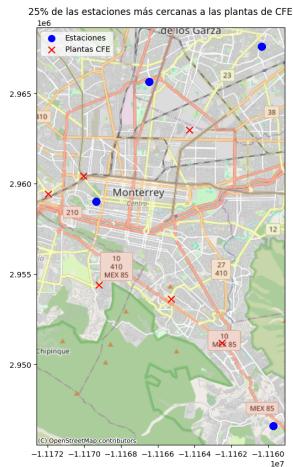
Para respaldar nuestro análisis visual, aplicamos un enfoque estadístico utilizando los métodos de KNN y PCA. Primero, identificamos el 25% de las estaciones de monitoreo más cercanas a cada planta de producción (CFE, Ternium y Cemex) a partir de un conjunto de n puntos disponibles. Utilizando KNN (Vecinos Más Cercanos), determinamos qué estaciones se encuentran geográficamente más próximas a cada planta, asegurando que los datos utilizados fueran los más representativos de las áreas cercanas a las fuentes de emisión.

Con esta selección de estaciones, procedimos a realizar un Análisis de Componentes Principales (PCA).

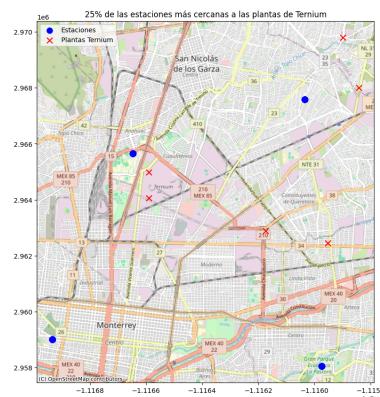
Las imágenes 4, 5 y 6 ilustran las estaciones seleccionadas como las más cercanas para cada planta (CFE, Ternium y Cemex). Este proceso de selección y análisis estadístico nos

proporcionó una base sólida para entender la relación entre las plantas de producción y la distribución de contaminantes en sus alrededores

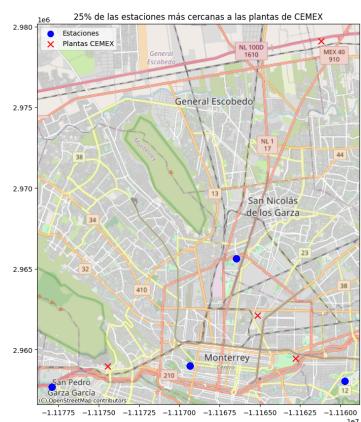
1) *Imagen 4. 25% de las estaciones más cercanas a las plantas de CFE*



2) *Imagen 5 .25% de las estaciones más cercanas a las plantas de Ternium*



3) *Imagen 6 .25% de las estaciones más cercanas a las plantas de Cemex*



Después, en base a las localizaciones, realizamos el análisis de componentes principales.

La separación primero con los KNN y posteriormente como encontramos el 25% más cercano de cada estación con cada fábrica.

6. Análisis

6.1. PCA, KNN y contaminantes más evidentes por localización

6.1.1. Aplicación KNN

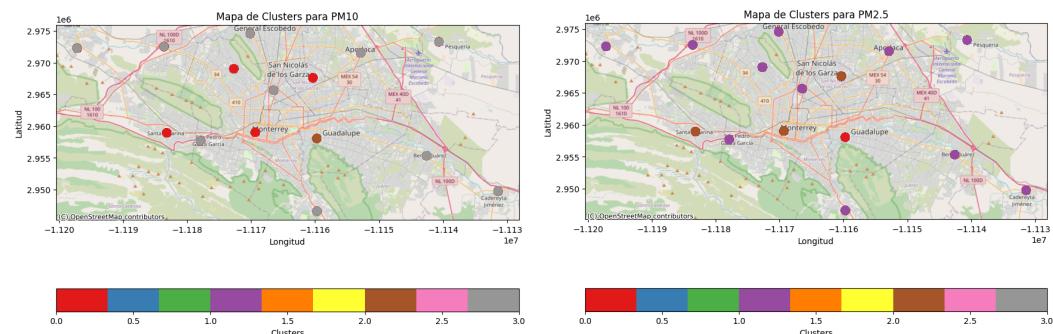


Imagen 7. Mapa de clústeres para PM10 y PM2.5

Para evaluar la selección de las estaciones de monitoreo de calidad del aire, utilizamos el algoritmo de KNN (K-Nearest Neighbors). Este método permitió agrupar las plantas industriales en función de los contaminantes que las afectan más significativamente. Los contaminantes analizados fueron CO, NOX, PM10, PM2.5, y SO2.

El algoritmo KNN identificó similitudes claras entre las estaciones de monitoreo más afectadas por ciertos contaminantes y las plantas industriales cercanas previamente identificadas en el análisis. Esto significa que el agrupamiento realizado por KNN respaldó y validó la selección inicial de estaciones de monitoreo, asegurando que las plantas y las estaciones se correspondan en términos de impacto de contaminantes.

Entre los contaminantes analizados, PM10, PM2.5, y SO2 se destacaron como los más relevantes para determinar la similitud entre las estaciones de monitoreo y las plantas cercanas. Esto refuerza la idea de que estos tres contaminantes tienen un impacto predominante en la calidad del aire de la zona.

Se confirmó que los contaminantes PM10, PM2.5 y SO2 son los factores principales que afectan tanto a las plantas industriales como a las estaciones cercanas. Esto demuestra la solidez del análisis y garantiza que las estaciones elegidas son representativas de las emisiones industriales, facilitando un monitoreo más preciso y adecuado de la calidad del aire en la región.

6.1.2. Aplicación PCA

Tabla 3. Resultados PCA.

Planta	Componente Principal	Contaminantes Dominantes	Variabilidad
CFE	CP1	PM10, PM2.5, NOx, CO	40%
	CP2	CO, SO2	20%
TERNIUM	CP1	PM10, PM2.5, NOx	40%
	CP2	CO, SO2	20%
CEMEX	CP1	PM10, PM2.5, NOx	40%
	CP2	CO, SO2	20%

Elegimos el Análisis de Componentes Principales porque nos ayuda a reducir la cantidad de variables sin perder información importante. Esto hace que sea más fácil analizar los contaminantes, ya que nos permite identificar patrones clave y eliminar cosas que se repiten, lo cual hace que el análisis sea más eficiente.

En CFE y Ternium, el primer componente principal (CP1) está principalmente influenciado por PM10, PM2.5, NOX y CO, lo que explica alrededor del 40% de la variabilidad en las emisiones. Esto significa que estos contaminantes son los que más afectan las variaciones en las emisiones de estas plantas.

El segundo componente principal (CP2) resalta la variabilidad de CO y SO2, capturando aproximadamente el 20% de la variabilidad. Aunque estos contaminantes no son tan influyentes como los del primer componente, siguen siendo importantes para entender cómo cambian las emisiones en CFE y Ternium.

En Cemex, el primer componente principal también está influenciado por PM10, PM2.5 y NOX, pero no tanto por CO. Esto nos dice que las emisiones de CO no tienen un impacto tan grande en la variabilidad de las emisiones de esta planta.

Esta diferencia sugiere que el perfil de contaminación de Cemex es distinto, probablemente por las características específicas de sus procesos de producción en la industria cementera.

6.2. Mapas de degradados e interpolación de valores de las estaciones

Esto se encuentra en: [SCRIPT5_GRAFICACIONVIENTO.ipynb](#)

6.2.1. Librería SCIPY

Características Técnicas:

- Se emplea interpolación cúbica mediante la función `griddata` de SciPy para generar superficies suaves entre los datos.
- Se crean mapas de calor utilizando GeoPandas, lo que facilita la visualización de la distribución de contaminantes.

Para llevar a cabo la interpolación, se utiliza la función `griddata` de la biblioteca SciPy, que permite crear superficies suavizadas a partir de datos conocidos. SciPy es una herramienta poderosa para la interpolación, ya que ofrece métodos avanzados que estiman valores en ubicaciones donde no se dispone de datos directos, como ocurre con los niveles de contaminación entre estaciones de monitoreo.

6.2.2. Interpolación cúbica

Nuestro código se basa en la función `griddata` de `scipy.interpolate` y utiliza el módulo `Transformer` de `pyproj`. Esta función opera bajo la siguiente fórmula:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3$$

La interpolación cúbica es una técnica que ajusta una función polinómica de tercer grado entre puntos de datos conocidos. Esta metodología resulta especialmente útil para modelar variaciones suaves y naturales en los datos de contaminación, evitando cambios bruscos que no son realistas.

En este caso, el algoritmo emplea los valores conocidos de contaminantes (como CO) junto con las condiciones del viento (velocidad y dirección) para inferir la distribución de los niveles de contaminación en áreas intermedias. Esto permite crear una malla de estimaciones que refleja de manera más precisa la calidad del aire en ubicaciones donde no se cuentan con sensores.

6.2.3. Resultados

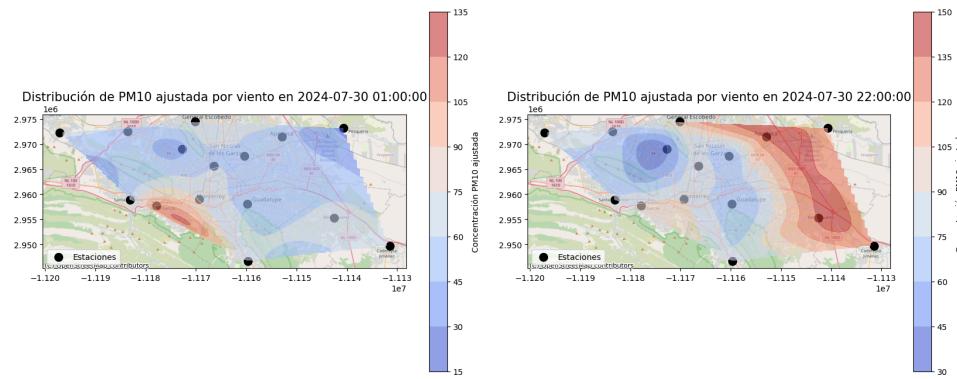


Imagen 9- Distribución de contaminante de PM10 en Monterrey , el 30 de julio de 2024 de 1am a 10pm.

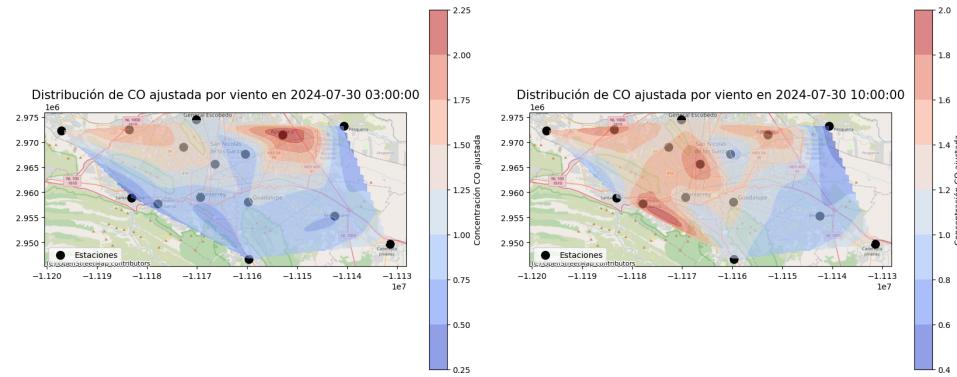


Imagen 10.- Distribución de contaminante de CO en Monterrey , el 30 de julio de 2024 de 3am a 10 am.

La distribución de monóxido de carbono (CO) y PM10 varía según las actividades humanas e industriales. Los niveles de CO aumentan durante las horas pico de la mañana y la tarde, cuando el tráfico vehicular es más intenso, y disminuyen en las horas de menor tránsito, lo que indica una relación directa entre el tráfico y las concentraciones de CO. En contraste, los niveles de PM10 reflejan la actividad de las fábricas, con mayores concentraciones durante las horas laborales y reducciones durante los períodos de inactividad industrial, como fines de semana. Esto sugiere que, para mejorar la calidad del aire, es crucial gestionar tanto el tráfico como las emisiones industriales en los momentos clave del día.

7. Conclusiones

El análisis de correlación geográfica reveló una relación significativa entre la ubicación de las fuentes de emisión y los niveles de contaminantes, destacando la influencia de los patrones industriales y ambientales en la distribución de la contaminación. Además, el manejo de datos mediante MICE con Regresión Bayesiana permitió imputar datos faltantes de manera eficiente, produciendo estimaciones robustas que mejoran la calidad del análisis. Este enfoque no solo facilitó una visión más precisa de los niveles de contaminación, sino que también permitió la comparación entre los reportes empresariales y las predicciones del modelo. Esto es crucial para identificar posibles inconsistencias en la información reportada, lo cual puede ayudar a mejorar la transparencia y la efectividad de las medidas de control de la contaminación en la región. En conjunto, estos métodos ofrecen una herramienta poderosa para el análisis ambiental y la toma de decisiones basadas en datos sólidos.

Así se proporciona un enfoque integral para abordar desafíos complejos en la gestión ambiental. Al identificar patrones espaciales relevantes y con la mejora en la calidad de los datos, se permite desarrollar estrategias más informadas y precisas para la mitigación de la contaminación. Además, la comparación entre los datos empresariales reportados y las predicciones del modelo no solo es útil para revelar posibles inconsistencias, sino que también así se fomenta la transparencia y la rendición de cuentas. Este enfoque integrador fortalece la capacidad de los tomadores de decisiones para diseñar políticas públicas más efectivas y sostenibles, promoviendo una gobernanza ambiental basada en la evidencia y adaptada a las realidades locales, como lo es en este caso el Sistema Integral de Monitoreo Ambiental (SIMA) de la Zona Metropolitana de Monterrey.

8. Referencias

Azur, M., & Stuart, E. (2011). *Multiple imputation by chained equations: What is it and how does*

it work? National Library Of Medicine.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/>

Centro de Calidad Ambiental. (2002). *Calidad Ambiental*, 7(4), Julio/Agosto. Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM).

<http://hdl.handle.net/11285/573525>

Sokhi, R. S., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., et al. (2022). Advances in air quality research – current and emerging challenges. *Atmospheric Chemistry and Physics*, 22(7), 4615–4703. <https://doi.org/10.5194/acp-22-4615-2022>

Universidad de Buenos Aires. (2008). *Contaminación atmosférica*. Facultad De Agronomía.

<https://www.agro.uba.ar/users/semmarti/Atmosfera/contatmosf.pdf>

World Health Organization: WHO. (2019, July 30). *Air pollution*.

https://www.who.int/health-topics/air-pollution#tab=tab_1

9. Anexos

Manejo de la información

Limpieza: [SCRIPT1_LIMPIEZA.ipynb](#)

Exploración: [SCRIPT2_EXPLORACION.ipynb](#)

Pruebas Estadísticas:

Interdependencia:

PCA: [SCRIPT3_PRUEBAS.ipynb](#)

KNN: [SCRIPT4_CLUSTERS.ipynb](#)

Dependencia lineal:

Regresiones y tests: [SCRIPT3_PRUEBAS.ipynb](#)

Regresion multiple: [VERIFICACION.Rmd](#)

Manejo GeoEspacial:

Interpolación:

[SCRIPT5_GRAFICACIONVIENTO.ipynb](#)

Agrupación estaciones e industrias:

[SCRIPT4_CLUSTERS.ipynb](#)

[SCRIPT3_PRUEBAS.ipynb](#)

Bases de datos:

[Carpeta a base de datos](#)

Todo:

 SIMA